Web scraping |Easy Steps for Web Scraping | Methods | Python

Web scraping itself is not illegal. As a matter of fact, web scraping or web crawling were historically associated with well-known search engines like Google or Bing. These search engines crawl sites and index the web.

Mr. Viraj Shelar Follow
Sep 4, 2020 · 4 min read

Web Scraping

Beginning with proccess of web scraping requires following :

> ***Required Imports :***

1. xlrd

xlrd is a module that allows Python to read data from Excel files.

Installation:

pip install xlrd

2. Selenium

Selenium WebDriver is one of the most popular tools for Web UI Automation.

Installation:

pip install selenium

3. Beautiful Soup

Beautiful Soup is a library that makes it easy to scrape information from web pages

Installation:

pip install beautifulsoup4

4. Pandas

Pandas is a Python package that provides fast, flexible, and expressive data structures designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time-series data both easy and intuitive.

Installation:

pip install pandas

5. x**lwt**

xlwt is a library for developers to use to generate spreadsheet files compatible with Microsoft Excel versions 95 to 2003.

Installation:

pip install xlwt

In the following steps, we are doing the Web Scraping for fetching the details of the Organization.

> *Step 1: Importing the imports*

```
import xlrd

from selenium import webdriver

from bs4 import BeautifulSoup

import pandas as pd

import xlwt

from xlwt import WorkbookStep 2: Importing the imports
```

> *Step 2: creating object for Excel Sheet*

```
wb1 = Workbook()
```

> *Step 3 : creating Sheet in Excel*

```
sheet1 = wb1.add_sheet('Sheet 1')
```

> *Step 4: Variables for Columns and Rows to handle the loop*

```
col=0

row=0
```

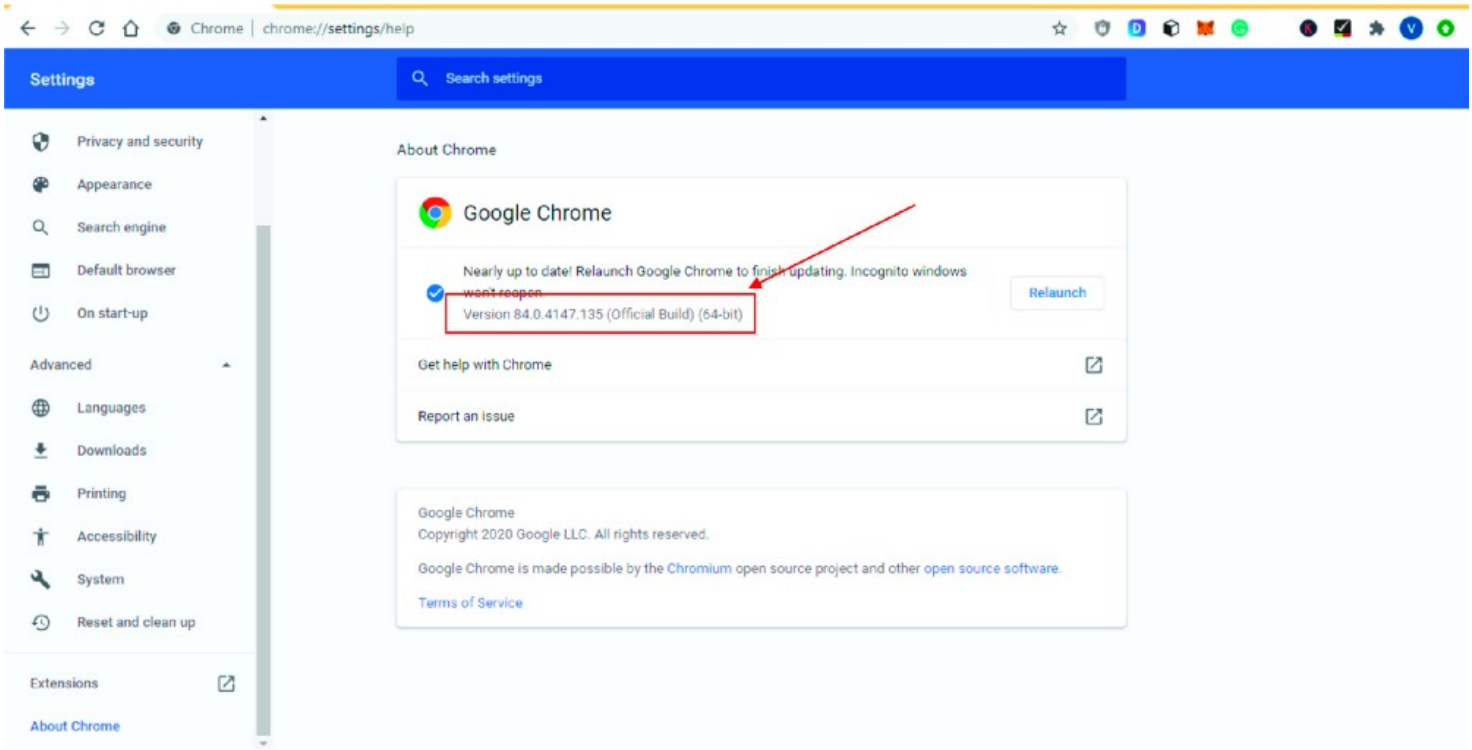
> *Step 5: Variable for Path of Excel Sheet containing the dataset to used for Scraping*

```
loc = ("C:/Users/Thinksprout
Infotech/Desktop/CompReg_31AUGUST2020.xlsx")
```

We are using the Test Chrome browser.

It is mandatory to have a chrome browser.

Check the your Chrome Version and download the same Compatable to yours.



Can be downloaded Chrome Test Web browser Drivers from :

> *Step 6: Variable for setting the test web browser drivers location*

```
driver = webdriver.Chrome("C:/Users/ Thinksprout Infotech
/Desktop/chromedriver_win32/chromedriver"
```

> *Step 7: Workbook object to read the Dataset from excel*

```
wb = xlrd.open_workbook(loc)

sheet = wb.sheet_by_index(0)
```

> *Step 8: traversing the Dataset*

Continue the Loop till end

```
for i in range(sheet.nrows):

      cin = sheet.cell_value(i, 0)

      name = sheet.cell_value(i, 1)
```

> *Step 9: URL that is to web scraped*

```
#HiddenData is the URL of website to be webscrapped
#brackets '{ }' represents the data for parameters to the website URL

stuff_in_string = "https://HiddenData/{}/{}".format(name,cin)
```

> *Step 10: Sending data of URL to the Web browser*

```
driver.get(stuff_in_string)

content = driver.page_source

soup = BeautifulSoup(content,"html.parser")
```

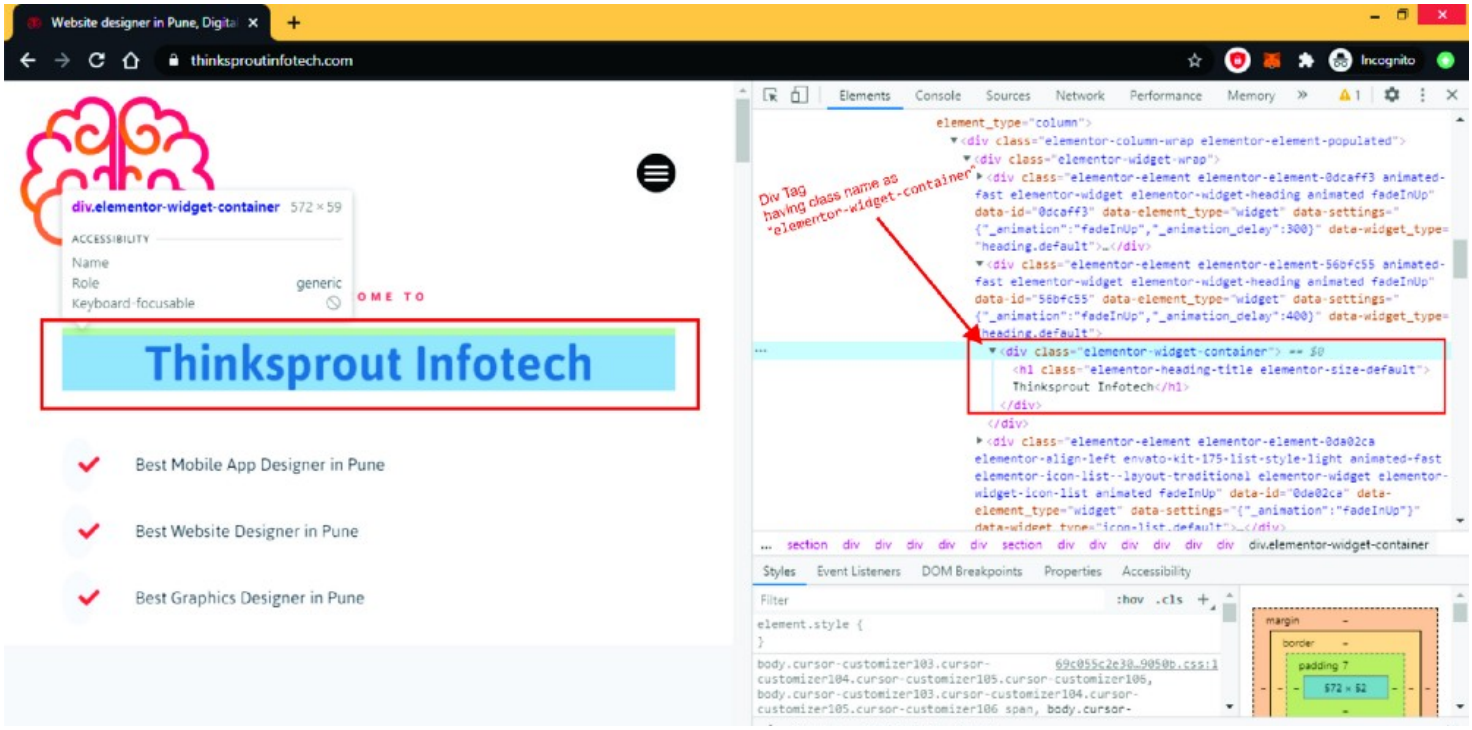
> *Step 11: Data for actual Website Tags that has to scraped and Data has to be fetched*

For e.g:

https://thinksproutinfotech.com

If we want to get the Thinksprout Infotech Name from my Organisation website URL

Right Click on the Webpage — > click inspect




```
content1 = soup.find('div', {"class": " elementor-widget-container"})
article = ''
```

Step 12: Exporting Data to New Excel-Sheet

```
#another Loop in our previous for loop
for i in content1.findAll('p'):
    article =i.text
    print(article)
    sheet1.write(row, col, article)
    wb1.save('C:/Users/Viraj Shelar/Desktop/data2.csv')
    row+=1
```

Step 13: Closing the Test Chrome Browser

```
#outside the second for loop i.e inside second first for loop
driver.quit()
```

This Loop will continue until the whole data from the dataset isn't parsed.

Conclusion:

After every successful execution of data from the dataset the Chrome Test Browser will restart itself for proceeding with Data values from data.

When sparsing the data and fetching the values simultaneously it will store the data fetched into the new Excel sheet. Simultaneously execution is done because storing of

data in Tuple, List or Dictionary might cause to Data loss due to failure in internet connection.

To avoid this Drawback simultaneous execution is preferred.

References:

1. https://www.geeksforgeeks.org/reading-excel-file-using-python/
2. https://www.geeksforgeeks.org/writing-excel-sheet-using-python/?ref=lbp
3. https://www.extendoffice.com/documents/excel/2517-excel-if-column-contains-value-text-then-copy-cell.html
4. https://matthew-brett.github.io/teaching/string_formatting.html
5. Image : https://unsplash.com/photos/RYyr-k3Ysqg